DOCUMENT RESUME

ED 397 061 TM 022 372

AUTHOR Bizot, Elizabeth B.; Goldman, Steven H.
TITLE The Practical Impact of IRT Models and Parameters When Converting a Test to Adaptive Format.
PUB DATE Apr 94
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability; *Adaptive Testing; *Computer Assisted Testing; *Estimation (Mathematics); High Schools; *High School Students; *Item Response Theory; Models; Selection; *Test Format; Vocabulary
IDENTIFIERS Calibration; Data Conversion; *Three Parameter Model; Two Parameter Model

ABSTRACT

A study was conducted to evaluate the effects of choice of item response theory (IRT) model, parameter calibration group, starting ability estimate, and stopping criterion on the conversion of an 80-item vocabulary test to computer adaptive format. Three parameter calibration groups were tested: (1) a group of 1,000 high school seniors, (2) a group of 1,000 high school freshmen, and (3) 300 of this second group retested as seniors. Two methods for setting the initial ability estimate, a random-based estimate and an ability-based estimate, were explored using two-parameter-logistic, three-parameter logistic with "c" parameter fixed at 0.2 (2.5 parameter), and full three-parameter logistic models. Alternatives were tested against a database of 2,697 people (including the calibration group) who had taken the full 80-item test. Results indicate that adaptive testing scores are relatively robust to differences in IRT models and parameters. The full three-parameter model was the best theoretical match to the test and gave the best practical results, but the 2.5 parameter model results were not much different. Five tables present analysis results. (Contains 3 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 397 061

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

ELIZABETH B. BIZOT

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

The Practical Impact of IRT Models and Parameters When Converting a Test to Adaptive Format

by

Elizabeth B. Bizot

Steven H. Goldman

Ball Foundation
800 Roosevelt Road
Suite C-120
Glen Ellyn, IL 60137

BEST COPY AVAILABLE

April, 1994

Paper Presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA

TM022372

Item Response Theory (IRT) provides the theoretical underpinning for converting a test from standard linear format to computer adaptive, but there are still questions about the implications of choices in some components of the testing process (Hambleton, Zaal, & Pieters, 1991). This study was conducted to evaluate the effects of choice of IRT model, parameter calibration group, starting ability estimate, and stopping criterion on the conversion of an 80-item vocabulary test to computer adaptive format.

Alternatives Evaluated

IRT Parameter Calibration Groups/Models

Item parameters are independent of the examinees used to estimate them as long as the parameter calibration group represents the full range of ability. We wanted to see the practical effect of a non-representative calibration group. Three groups were studied: L1, a group of 1000 high school seniors; L2, a group of 1000 high school freshmen (who scored much lower on the test than did L1); and L2 Retest (L2R), 300 of the original L2 group who were retested as high school seniors and who were in general higher achieving than L1.

Most adaptive testing applications use the three parameter model as the best theoretical fit to multiple choice data (Hambleton, Zaal, & Pieters, 1991), but there can

be difficulty in estimating the c parameter adequately (Hambleton, 1989) and so other models may actually perform better. Our intention was to estimate item parameters for each of the three groups using three models: 2 parameter logistic (2PL), 3-parameter logistic with c parameter fixed at .2 (2.5PL), and full 3-parameter logistic (3PL). However, BILOG was unable to calculate the 2.5PL model for the L2 freshman group, and so a total of eight model/group parameter estimates were evaluated. Tables 1, 2, and 3 display item parameters for each of the three groups for the three parameter model.

Insert Tables 1, 2, and 3 About Here

Initial Ability Estimate/Initial Item Selection

Hambleton, Zaal, and Pieters (1991) suggest that prior information on ability can improve testing efficiency by helping select a correct starting point, but that most researchers believe that starting with an item of moderate difficulty yields adequate test performance. We evaluated two methods for setting the initial ability estimate (from which items appropriate for that ability level would be chosen). For a moderate-difficulty start, we used a random ability estimate between -2 and +2, and for an ability-based estimate we used a formula based on years of previous education.

Stopping Criterion

Hambleton, Zaal, and Pieters (1991) report that most adaptive testing programs

stop testing based on the standard error, a preselected number of items, or some combination of both. This study investigated stopping at total test information equal to 7.5, 11.1, and 16 (equivalent to a standard-error based criterion because the standard error is a function of the test information), or when 25 items had been administered.

Method

Subjects

The alternatives were tested against a database of 2697 people who had taken the full 80-item test. The 2697 included people from the calibration samples (L1, L2 freshman, and L2 Retest), as well as career counseling clients and occupational research subjects. Only people who had responded to all 80 items were included; this selection eliminated some low-scoring individuals, but was necessary because responses to any item might be needed in simulating the adaptive test. Subjects were 51% female; the ethnic distribution was 14% African-American, 69% Caucasian, 15% Hispanic, and 2% other. Age ranges were 28% 13-15, 40% 16-18, 10% 19-25 and 22% 26 and over. Scores on the standard 80-item test ranged from 8 to 80 ($M = 41.61$, $SD = 21.33$).

Instrument

The test is an 80-item multiple choice vocabulary test which is part of the Ball Aptitude Battery (Ball Foundation, 1993). Although additional items are available, for purposes of this study, only the 80 items from Form A of the standard test were

included in the item pool. The test is sufficiently unidimensional for application of IRT; alpha reliability is .98, and factor analysis shows that the first factor accounts for 35% of the variance with no other large factors.

Procedure

A BASIC program was written to simulate adaptive administration of the test to each person in the database using the maximum-information method for selecting the next item to administer. Each person was "tested" 16 times: once for each starting estimate (random or education-based) for each of the eight item parameter groups. Ability estimates and number of items administered were recorded at each of the three information-based stopping points (if they were reached) and at the final 25 items for each administration. In addition, an estimate of the full 80-item score was computed by summing the probability of success on each item for a person of the given ability.

The results of each alternative for each decision were evaluated by examining the average number of items administered and the average absolute difference between the estimated and the actual 80-item score under each condition. In computing the averages, for cases where an information-based stopping criterion was not met, data for the 25-item administration for that person was used, as would happen during actual adaptive testing. Because alternatives might perform differently at different points on the ability scale, results were examined separately for four levels

based on scores on the original test: those scoring 0-20 (N = 529), 21-40 (N = 958), 41-60 (N = 512), and 61-80 (N = 698).

Results

IRT Model/Parameter Groups

As demonstrated in Table 4, average absolute difference scores ranged from 2.80 to 12.04, with most in the range of 3 - 6. Average number of items administered ranged from 7.2 to 25 (the latter implying that information-based stopping criteria were never met for that combination of alternatives).

Insert Table 4 About Here

As expected, the two parameter sets based on the L2 freshman group did not perform well, particularly in the upper ability ranges where the average difference scores were between 9 and 12. Even in the lower ability ranges these models tended to require more items administered.

The difference scores resulting from the L1 group (N = 1000) and the L2 Retest group (N = 300) were very similar, indicating that there was little practical effect of using the smaller and somewhat less diverse group to estimate the item parameters, even in the full three parameter model. Differences between the two groups in number of items administered varied across ability levels, but was never more than 2 items.

With both the L1 and L2 Retest parameters, the 2PL model required fewer items

at the low end but more items at the high end. On the other hand, the 2.5PL and 3PL models required more items at the low ability end (averaging around 22) but fewer at the high, especially in the 41-60 score range where 8 - 9 items were often sufficient.

Overall, taking into account both difference scores and the number of items administered, the full 3PL model based on the L1 group was most effective, but several of the other models also performed adequately.

Initial Ability Estimate

As seen in Table 5, using a random starting ability rather than an education-based starting point required, on average, about one additional item and never more than two. Thus, differences appear to be minimal between groups based on starting point used.

Insert Table 5 About Here

Stopping Criterion

The highest stopping criterion (total information = 16) was almost never met, resulting in administration of the full 25 items in almost all cases. Between the other two information-based stopping criteria (7.5 and 11.1), the more stringent criterion resulted in a lower average difference about one-half to one point, usually at a cost of administering about 5 to 7 additional items. In all cases, there is a smaller average difference when more items are administered.

Discussion

Overall, the results indicate that adaptive testing scores are relatively robust to the differences in IRT models and parameters. The similarities between results for the L1 (N = 1000) and L2 Retest (N = 300) groups demonstrate that although the item a, b, and c parameters were somewhat different in the two models, the practical impact on adaptive testing was minimal. The 3PL model, including guessing, is the best theoretical match to the test and also produces the best practical results, but results from the 2.5PL model are not much different, again indicating a practical robustness to theoretical differences.

Having prior information on which to base the starting ability estimate reduced the number of test items somewhat, and raising the total amount of information required before stopping the computer-administered test reduces the difference between the adaptive score and the full linear score, but again both differences are small in practical terms.

References

Ball Foundation (1993). Ball Aptitude Battery Technical Manual. Glen Ellyn, Il: Ball Foundation.

Hambleton, R. K., Zaal, J. N., & Pieters, J. M. P. (1991). Computerized adaptive testing: Theory, applications, and standards. In R. K. Hambleton & J. N. Zaal (Eds.) Advances in educational and psychological testing: Theory and applications (pp. 341-366).

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), Educational Measurement, 3rd Edition (pp. 147-200).

Table 1

Item Parameter Estimates for the Three Parameter Model
For Group L1 (N = 1000)

Item	Parameter A	Parameter B	Parameter C
01	1.239	-0.863	0.209
02	0.873	-1.809	0.108
03	1.013	-0.777	0.131
04	1.714	-0.567	0.129
05	1.173	-1.353	0.161
06	0.874	-1.660	0.108
07	1.329	0.159	0.126
08	2.415	-0.221	0.195
09	1.597	1.379	0.397
10	1.141	-0.741	0.093
11	0.909	-0.237	0.071
12	0.991	-0.074	0.065
13	0.973	-0.281	0.156
14	1.032	-1.212	0.110
15	1.397	0.169	0.082
16	1.059	0.457	0.117
17	0.644	0.150	0.131
18	1.237	0.214	0.151
19	0.528	-0.508	0.158
20	1.065	0.549	0.109
21	1.568	0.451	0.156
22	1.446	0.250	0.246
23	1.456	0.643	0.144
24	1.703	0.521	0.134
25	1.378	0.426	0.076
26	0.819	-0.335	0.060
27	0.929	0.195	0.075
28	1.015	0.441	0.097
29	1.432	0.359	0.135
30	1.459	0.570	0.207
31	0.700	0.980	0.213
32	0.857	0.921	0.166
33	0.793	0.642	0.049
34	0.987	0.819	0.194
35	0.657	0.069	0.149
36	1.075	0.631	0.109
37	0.592	-0.028	0.079
38	0.724	0.351	0.162
39	1.160	0.307	0.100
40	1.459	1.622	0.166
41	1.707	0.949	0.049
42	0.956	-0.208	0.126

Table 1 Cont.

Item	Parameter A	Parameter B	Parameter C
43	1.180	0.987	0.160
44	1.700	1.126	0.075
45	1.540	1.752	0.211
46	1.359	0.926	0.121
47	0.669	0.960	0.118
48	1.861	1.340	0.132
49	1.049	0.779	0.109
50	1.800	1.299	0.317
51	0.829	1.804	0.094
52	1.536	1.201	0.171
53	0.986	0.763	0.112
54	0.941	1.805	0.169
55	0.935	0.406	0.169
56	1.434	1.314	0.112
57	1.782	1.071	0.154
58	0.958	1.550	0.178
59	1.138	1.380	0.126
60	1.005	0.990	0.047
61	1.164	0.897	0.046
62	1.041	1.294	0.060
63	0.950	2.023	0.132
64	1.290	1.472	0.059
65	1.250	2.008	0.249
66	1.876	1.669	0.166
67	1.343	1.290	0.145
68	1.428	1.457	0.101
69	1.241	1.676	0.094
70	0.732	1.983	0.038
71	1.345	1.463	0.191
72	1.386	1.244	0.059
73	1.309	1.758	0.071
74	1.278	1.467	0.159
75	1.012	1.911	0.107
76	1.225	1.712	0.115
77	1.355	1.921	0.132
78	1.201	2.151	0.140
79	0.368	3.842	0.139
80	0.988	2.425	0.042

Table 2

Item Parameter Estimates for the Three Parameter Model
For Group L2 (N = 1000)

Item	Parameter A	Parameter B	Parameter C
01	1.389	0.101	0.264
02	1.016	-0.905	0.123
03	0.914	-0.087	0.083
04	1.447	0.044	0.198
05	1.096	-0.433	0.159
06	0.842	-0.981	0.120
07	1.085	0.890	0.065
08	1.636	0.783	0.169
09	0.415	1.642	0.246
10	1.247	0.577	0.187
11	0.928	1.138	0.103
12	0.910	0.587	0.085
13	0.913	0.551	0.134
14	1.141	-0.164	0.105
15	1.574	1.338	0.090
16	1.093	1.536	0.190
17	0.955	1.250	0.239
18	1.949	1.363	0.226
19	0.523	0.401	0.149
20	1.172	1.560	0.105
21	1.648	1.100	0.125
22	1.785	1.692	0.256
23	1.833	1.649	0.169
24	2.034	1.414	0.111
25	1.889	1.167	0.069
26	0.723	0.441	0.067
27	0.972	1.633	0.125
28	1.118	1.514	0.120
29	1.221	1.555	0.087
30	0.799	1.174	0.145
31	0.507	1.476	0.116
32	0.912	2.438	0.177
33	0.701	1.346	0.087
34	1.068	1.681	0.161
35	0.682	1.075	0.148
36	1.816	1.709	0.174
37	0.508	0.677	0.114
38	0.821	1.525	0.251
39	1.502	0.777	0.141
40	1.723	2.502	0.194
41	2.168	1.807	0.083
42	0.822	0.901	0.154

Table 2 cont.

Item	Parameter A	Parameter B	Parameter C
43	1.747	1.733	0.200
44	1.999	2.279	0.063
45	1.548	2.115	0.125
46	2.054	2.006	0.158
47	0.873	2.240	0.159
48	1.859	2.343	0.153
49	0.929	1.929	0.182
50	0.712	2.302	0.283
51	0.535	3.164	0.090
52	1.384	2.070	0.102
53	1.184	1.782	0.160
54	1.141	3.018	0.185
55	0.761	1.095,	0.157
56	1.438	2.269	0.112
57	1.415	1.988	0.108
58	1.618	2.075	0.174
59	1.267	2.162	0.157
60	1.235	2.194	0.079
61	1.248	1.868	0.062
62	1.236	2.556	0.101
63	2.759	2.241	0.200
64	1.431	2.711	0.081
65	0.794	3.489	0.186
66	1.316	3.085	0.179
67	1.261	2.876	0.158
68	1.321	2.023	0.107
69	0.998	2.038	0.081
70	1.252	2.499	0.097
71	2.099	2.327	0.193
72	1.413	2.546	0.062
73	1.562	2.732	0.101
74	0.891	2.422	0.142
75	0.959	2.512	0.109
76	1.019	2.386	0.100
77	1.262	3.242	0.132
78	0.815	3.218	0.116
79	0.699	2.868	0.153
80	2.144	2.921	0.071

Table 3

Item Parameter Estimates for the Three Parameter Model
For Group L2R (N = 300)

Item	Parameter A	Parameter B	Parameter C
01	1.416	-1.208	0.137
02	0.793	-2.373	0.148
03	1.144	-1.068	0.198
04	1.034	-1.245	0.133
05	1.490	-1.356	0.158
06	1.152	-1.756	0.178
07	1.376	-0.060	0.152
08	2.046	-0.394	0.135
09	1.116	0.758	0.281
10	1.386	-0.811	0.220
11	0.621	-0.665	0.130
12	0.782	-0.258	0.157
13	1.059	-0.275	0.288
14	1.091	-0.910	0.147
15	1.714	0.425	0.166
16	1.091	0.318	0.136
17	0.437	-0.314	0.176
18	1.019	-0.090	0.095
19	0.423	-1.058	0.179
20	1.123	0.170	0.163
21	1.264	0.164	0.107
22	1.952	0.075	0.153
23	1.553	0.456	0.112
24	1.507	0.497	0.117
25	1.389	0.425	0.104
26	0.492	-0.832	0.132
27	1.032	0.264	0.107
28	0.951	-0.042	0.119
29	1.066	0.594	0.157
30	1.195	0.357	0.220
31	0.460	0.884	0.229
32	0.750	1.255	0.213
33	0.993	0.847	0.113
34	1.021	0.271	0.172
35	0.502	-0.328	0.134
36	0.864	0.769	0.143
37	0.521	-0.381	0.175
38	0.602	0.379	0.185
39	2.263	-0.288	0.176
40	0.617	1.485	0.206
41	1.358	1.082	0.055
42	0.771	-0.671	0.183

Table 3 Cont.

Item	Parameter A	Parameter B	Parameter C
43	1.295	0.616	0.182
44	1.287	1.198	0.062
45	1.407	1.680	0.161
46	1.082	0.914	0.093
47	0.731	1.209	0.157
48	1.550	1.037	0.116
49	0.628	-0.304	0.137
50	1.319	1.299	0.334
51	0.985	2.000	0.168
52	0.800	1.258	0.107
53	0.896	0.398	0.108
54	1.310	1.944	0.168
55	0.677	0.025	0.202
56	1.132	0.969	0.076
57	1.213	1.038	0.108
58	0.648	0.951	0.144
59	0.896	1.409	0.152
60	0.898	0.771	0.059
61	1.517	0.857	0.103
62	1.080	1.093	0.101
63	1.309	1.635	0.129
64	1.483	1.092	0.092
65	1.392	1.777	0.292
66	1.784	1.626	0.147
67	0.941	1.415	0.140
68	1.223	1.732	0.112
69	1.006	1.490	0.094
70	1.188	1.803	0.080
71	1.337	0.862	0.206
72	1.073	1.463	0.071
73	1.355	1.961	0.125
74	1.352	1.443	0.201
75	1.279	1.545	0.151
76	1.520	1.595	0.118
77	0.700	2.672	0.150
78	0.707	2.165	0.165
79	0.973	2.724	0.221
80	1.208	1.497	0.046

Table 4

Average Absolute Difference and Average Items Administered by Group and Vocabulary Score with Educational Level Starting Ability

VO SCORE		1 (0-20)		2 (21-40)		3 (41-60)		4 (61-80)	
CUTOFF		A	B	A	B	A	B	A	B
L1/2	MDIFF	4.17	3.57	4.67	3.81	5.03	3.88	2.89	2.85
	MITEM	9.90	17.30	8.40	14.60	17.40	23.40	24.90	25.00
L1/2.5	MDIFF	4.09	3.71	4.37	3.70	7.10	6.38	4.21	3.53
	MITEM	23.90	24.80	10.70	16.90	7.20	9.70	12.30	14.40
L1/3	MDIFF	3.07	2.70	4.96	4.13	6.32	5.80	3.41	2.95
	MITEM	21.40	24.40	9.80	15.90	8.00	11.40	15.70	18.80
L2/2	MDIFF	2.93	2.79	3.35	3.20	.4.30	4.30	11.96	11.96
	MITEM	21.60	25.00	22.10	25.00	25.00	25.00	25.00	25.00
L2/3	MDIFF	2.60	2.39	4.48	3.95	6.83	5.86	9.76	6.86
	MITEM	22.90	24.80	11.40	17.00	9.10	11.2	14.60	17.00
L2R/2	MDIFF	4.21	3.56	4.74	3.90	4.03	3.31	2.80	2.80
	MITEM	9.20	15.90	8.10	14.00	18.20	24.10	24.90	25.00
L2R/2.5	MDIFF	4.70	4.15	5.33	4.55	5.79	5.13	4.23	3.29
	MITEM	23.40	24.90	11.10	17.80	7.90	11.30	14.10	16.40
L2R/3	MDIFF	3.54	2.80	5.26	4.29	5.99	5.06	3.15	2.90
	MITEM	22.30	24.70	9.90	16.40	8.80	12.80	16.70	19.70

Note. MDIFF = Mean absolute difference scores
MITEM = Mean of items administered
A = 7.5 stopping criterion total
B = 11.1 stopping criterion total
VO SCORE = Vocabulary score range
LR/2 = Group L2R at 2 parameter model
L1/2 = Group L1 at 2 parameter model
L1/2.5 = Group L1 at 2.5 parameter model
L1/3 = Group L1 at 3 parameter model
L2/2 = Group L2 at 2 parameter model
L2/3 = Group L2 at 3 parameter model
L2R/2.5 = Group L2R at 2.5 parameter model
L2R/3 = Group L2R at 3 parameter model

Table 5

Average Absolute Difference and Average Items Administered by Group and Vocabulary Score with Random Starting Ability

VO SCORE		1 (0-20)		2 (21-40)		3 (41-60)		4 (61-80)	
CUTOFF		A	B	A	B	A	B	A	B
L1/2	MDIFF	4.18	3.65	4.57	3.80	4.83	3.88	2.92	2.88
	MITEM	10.90	17.80	9.30	15.10	17.90	23.60	24.90	25.00
L1/2.5	MDIFF	4.43	3.97	5.88	4.70	7.34	7.02	4.30	3.33
	MITEM	23.10	24.70	12.40	18.10	7.90	10.30	12.60	14.70
L1/3	MDIFF	3.32	2.95	5.50	4.58	6.42	5.80	3.43	2.88
	MITEM	22.10	24.60	10.90	16.70	8.60	12.00	16.10	19.20
L2/2	MDIFF	2.94	2.77	3.35	3.20	4.35	4.35	12.04	12.04
	MITEM	21.60	25.00	22.20	25.00	24.99	25.00	25.00	25.00
L2/3	MDIFF	3.15	2.86	5.03	4.40	7.04	5.91	9.40	6.75
	MITEM	22.90	24.70	11.90	17.30	8.90	11.13	15.10	17.40
L2R/2	MDIFF	4.45	3.64	4.74	3.92	4.04	3.36	2.88	2.88
	MITEM	10.20	16.50	8.90	14.60	18.60	24.20	24.90	25.00
L2R/2.5	MDIFF	4.56	4.18	5.87	5.04	6.92	6.34	4.11	3.35
	MITEM	23.80	24.90	12,80	18.90	9.20	12.60	14.60	16.90
L2R/3	MDIFF	3.68	2.88	5.59	4.63	6.61	5.47	3.31	2.91
	MITEM	22.80	24.70	11.30	17.30	9.90	13.90	17.70	20.10

Note. MDIFF = Mean absolute difference scores
MITEM = Mean of items administered
A = 7.5 stopping criterion total
B = 11.1 stopping criterion total
VO SCORE = Vocabulary score range
L1/2 = Group L1 at 2 parameter model
L1/2 = Group L1 at 2 parameter model
L1/2.5 = Group L1 at 2.5 parameter model
L1/3 = Group L1 at 3 parameter model
L2/2 = Group L2 at 2 parameter model
L2/3 = Group L2 at 3 parameter model
L2R/2.5 = Group L2R at 2.5 parameter model
L2R/3 = Group L2R at 3 parameter model